Near–Optimal Bayesian Active Learning
with Noisy Observations

Daniel Golovin
Caltech

Andreas Krause
Caltech

Debajyoti Ray
Caltech

Abstract

We tackle the fundamental problem of Bayesian active learning with noise, where we need to adaptively select from a number of expensive tests in order to identify an unknown hypothesis sampled from a known prior distribution. In the case of noise–free observations, a greedy algorithm called generalized binary search (GBS) is known to perform near–optimally. We show that if the observations are noisy, perhaps surprisingly, GBS can perform very poorly. We develop EC^2, a novel, greedy active learning algorithm and prove that it is competitive with the optimal policy, thus obtaining the first competitiveness guarantees for Bayesian active learning with noisy observations. Our bounds rely on a recently discovered diminishing returns property called adaptive submodularity, generalizing the classical notion of submodular set functions to adaptive policies. Our results hold even if the tests have non–uniform cost and their noise is correlated. We also propose EFFECX-TIVE, a particularly fast approximation of EC^2, and evaluate it on a Bayesian experimental design problem involving human subjects, intended to tease apart competing economic theories of how people make decisions under uncertainty.

1   Introduction

How should we perform experiments to determine the most accurate scientific theory among competing candidates, or choose among expensive medical procedures to accurately determine a patient's condition, or select which labels to obtain in order to determine the hypothesis that minimizes generalization error? In all these applications, we have to sequentially select among a set of noisy, expensive observations (outcomes of experiments, medical tests, expert labels) in order to determine which hypothesis (theory, diagnosis, classifier) is most accurate. This fundamental problem has been studied in a number of areas, including statistics [16], decision theory [12], machine learning [18, 7] and others. One way to formalize such active learning problems is *Bayesian experimental design* [6], where one assumes a prior on the hypotheses, as well as probabilistic assumptions on the outcomes of tests. The goal then is to determine the correct hypothesis while minimizing the cost of the experimentation. Unfortunately, finding this optimal policy is not just NP-hard, but also hard to approximate [5]. Several heuristic approaches have been proposed that perform well in some applications, but do not carry theoretical guarantees (e.g., [17]). In the case where observations are *noise-free*[1], a simple algorithm, *generalized binary search*[2] *(GBS)* run on a modified prior, is guaranteed to be competitive with the optimal policy; the expected number of queries is a factor of $O(\log n)$ (where n is the number of hypotheses) more than that of the optimal policy [14], which matches lower bounds up to constant factors [5].

The important case of *noisy* observations, however, as present in most applications, is much less well understood. While there are some recent positive results in understanding the label complexity of noisy active learning [18, 1], to our knowledge, so far there are no algorithms that are provably

---

[1]This case is known as the *Optimal Decision Tree* (ODT) problem.

[2]GBS greedily selects tests to maximize, in expectation over the test outcomes, the prior probability mass of eliminated hypotheses (i.e., those with zero posterior probability, computed w.r.t. the observed test outcomes).

competitive with the optimal sequential policy, except in very restricted settings [15]. In this paper, we introduce a general formulation of Bayesian active learning with noisy observations that we call the *Equivalence Class Determination* problem. We show that, perhaps surprisingly, generalized binary search performs poorly in this setting, as do greedily (myopically) maximizing the information gain (measured w.r.t. the distribution on equivalence classes) or the decision-theoretic value of information. This motivates us to introduce a novel active learning criterion, and use it to develop a greedy active learning algorithm called the **E**quivalence **C**lass **E**dge **C**utting algorithm (EC^2), whose expected cost is competitive to that of the optimal policy. Our key insight is that our new objective function satisfies *adaptive submodularity* [9], a natural diminishing returns property that generalizes the classical notion of submodularity to adaptive policies. Our results also allow us to relax the common assumption that the outcomes of the tests are conditionally independent given the true hypothesis. We also develop the **Eff**icient **E**dge **C**utting appro**X**imate objec**tive** algorithm (EFFECXTIVE), an efficient approximation to EC^2, and evaluate it on a Bayesian experimental design problem intended to tease apart competing theories on how people make decisions under uncertainty, including Expected Value [21], Prospect Theory [13], Mean-Variance-Skewness [11] and Constant Relative Risk Aversion [19]. In our experiments, EFFECXTIVE typically outperforms existing experimental design criteria such as information gain, uncertainty sampling, GBS, and decision-theoretic value of information. Our results from human subject experiments further reveal that EFFECXTIVE can be used as a real-time tool to classify people according to the economic theory that best describes their behaviour in financial decision-making, and reveal some interesting heterogeneity in the population.

2 Bayesian Active Learning in the Noiseless Case

In the Bayesian active learning problem, we would like to distinguish among a given set of hypotheses $\mathcal{H} = \{h_1, \ldots, h_n\}$ by performing tests from a set $\mathcal{T} = \{1, \ldots, N\}$ of possible tests. Running test t incurs a cost of $c(t)$ and produces an outcome from a finite set of outcomes $\mathcal{X} = \{1, 2, \ldots, \ell\}$. We let H denote the random variable which equals the true hypothesis, and model the outcome of each test t by a random variable X_t taking values in $\mathcal{X}$. We denote the observed outcome of test t by x_t. We further suppose we have a prior distribution P modeling our assumptions on the joint probability $P(H, X_1, \ldots, X_N)$ over the hypotheses and test outcomes. In the noiseless case, we assume that the outcome of each test is deterministic given the true hypothesis, i.e., for each $h \in \mathcal{H}$, $P(X_1, \ldots, X_N \mid H = h)$ is a deterministic distribution. Thus, each hypothesis h is associated with a particular vector of test outcomes. We assume, w.l.o.g., that no two hypotheses lead to the same outcomes for all tests. Thus, if we perform all tests, we can uniquely determine the true hypothesis. However in most applications we will wish to avoid performing every possible test, as this is prohibitively expensive. Our goal is to find an adaptive policy for running tests that allows us to determine the value of H while minimizing the cost of the tests performed. Formally, a policy π (also called a conditional plan) is a partial mapping π from partial observation vectors $\mathbf{x}_\mathcal{A}$ to tests, specifying which test to run next (or whether we should stop testing) for any observation vector $\mathbf{x}_\mathcal{A}$. Hereby, $\mathbf{x}_\mathcal{A} \in \mathcal{X}^\mathcal{A}$ is a vector of outcomes indexed by a set of tests $\mathcal{A} \subseteq \mathcal{T}$ that we have performed so far [3] (e.g., the set of labeled examples in active learning, or outcomes of a set of medical tests that we ran). After having made observations $\mathbf{x}_\mathcal{A}$, we can rule out inconsistent hypotheses. We denote the set of hypotheses consistent with event Λ (often called the *version space* associated with Λ) by $\mathcal{V}(\Lambda) := \{h \in \mathcal{H} : P(h \mid \Lambda) > 0\}$. We call a policy *feasible* if it is guaranteed to uniquely determine the correct hypothesis. That is, upon termination with observation $\mathbf{x}_\mathcal{A}$, it must hold that $|\mathcal{V}(\mathbf{x}_\mathcal{A})| = 1$. We can define the expected cost of a policy π by

$$c(\pi) := \sum_h P(h) c(\mathcal{T}(\pi, h))$$

where $\mathcal{T}(\pi, h) \subseteq \mathcal{T}$ is the set of tests run by policy π in case $H = h$. Our goal is to find a feasible policy π^* of minimum expected cost, i.e.,

$$\pi^* = \arg\min \{c(\pi) : \pi \text{ is feasible}\} \tag{2.1}$$

A policy π can be naturally represented as a decision tree T^π, and thus problem (2.1) is often called the *Optimal Decision Tree* (ODT) problem.

Unfortunately, obtaining an approximate policy π for which $c(\pi) \leq c(\pi^*) \cdot o(\log(n))$ is NP-hard [5]. Hence, various heuristics are employed to solve the Optimal Decision Tree problem and its variants.

---

[3]Formally we also require that $(x_t)_{t \in \mathcal{B}} \in \mathrm{dom}(\pi)$ and $\mathcal{A} \subseteq \mathcal{B}$, implies $(x_t)_{t \in \mathcal{A}} \in \mathrm{dom}(\pi)$ (c.f., [9]).

Two of the most popular heuristics are to select tests greedily to maximize the *information gain* (IG) conditioned on previous test outcomes, and *generalized binary search* (GBS). Both heuristics are greedy, and after having made observations $\mathbf{x}_\mathcal{A}$ will select

$$t^* = \arg\max_{t \in \mathcal{T}} \Delta_{\text{Alg}}(t \mid \mathbf{x}_\mathcal{A}) / c(t),$$

where $\text{Alg} \in \{\text{IG}, \text{GBS}\}$. Here, $\Delta_{\text{IG}}(t \mid \mathbf{x}_\mathcal{A}) := \mathbb{H}(\mathbf{X}_\mathcal{T} \mid \mathbf{x}_\mathcal{A}) - \mathbb{E}_{x_t \sim X_t \mid \mathbf{x}_\mathcal{A}}[\mathbb{H}(\mathbf{X}_\mathcal{T} \mid \mathbf{x}_\mathcal{A}, x_t)]$ is the marginal information gain measured with respect to the Shannon entropy $\mathbb{H}(\mathbf{X}) := \mathbb{E}_\mathbf{x}[-\log_2 P(\mathbf{x})]$, and $\Delta_{\text{GBS}}(t \mid \mathbf{x}_\mathcal{A}) := P(\mathcal{V}(\mathbf{x}_\mathcal{A})) - \sum_{x \in \mathcal{X}} P(X_t = x \mid \mathbf{x}_\mathcal{A}) P(\mathcal{V}(\mathbf{x}_\mathcal{A}, X_t = x))$ is the expected reduction in version space probability mass. Thus, both heuristics greedily chooses the test that maximizes the benefit-cost ratio, measured with respect to their particular benefit functions. They stop after running a set of tests $\mathcal{A}$ such that $|\mathcal{V}(\mathbf{x}_\mathcal{A})| = 1$, i.e., once the true hypothesis has been uniquely determined.

It turns out that for the (noiseless) Optimal Decision Tree problem, these two heuristics are equivalent [22], as can be proved using the chain rule of entropy. Interestingly, despite its myopic nature GBS has been shown [14, 7, 10, 9] to obtain near-optimal expected cost: the strongest known bound is $c(\pi_{GBS}) \leq c(\pi^*)(\ln(1/p_{\min}) + 1)$ where $p_{\min} := \min_{h \in \mathcal{H}} P(h)$. Let $\mathbf{x}_S(h)$ be the unique vector $\mathbf{x}_S \in \mathcal{X}^S$ such that $P(\mathbf{x}_S \mid h) = 1$. The result above is proved by exploiting the fact that $f_{GBS}(S, h) := 1 - P(\mathcal{V}(\mathbf{x}_S(h))) + P(h)$ is *adaptive submodular* and *strongly adaptively monotone* [9]. Call $\mathbf{x}_\mathcal{A}$ a *subvector* of $\mathbf{x}_\mathcal{B}$ if $\mathcal{A} \subseteq \mathcal{B}$ and $P(\mathbf{x}_\mathcal{B} \mid \mathbf{x}_\mathcal{A}) > 0$. In this case we write $\mathbf{x}_\mathcal{A} \prec \mathbf{x}_\mathcal{B}$. A function $f : 2^\mathcal{T} \times \mathcal{H}$ is called adaptive submodular w.r.t. a distribution P, if for any $\mathbf{x}_\mathcal{A} \prec \mathbf{x}_\mathcal{B}$ and any test t it holds that $\Delta(t \mid \mathbf{x}_\mathcal{A}) \geq \Delta(t \mid \mathbf{x}_\mathcal{B})$, where

$$\Delta(t \mid \mathbf{x}_\mathcal{A}) := \mathbb{E}_H[f(\mathcal{A} \cup \{t\}, H) - f(\mathcal{A}, H) \mid \mathbf{x}_\mathcal{A}].$$

Thus, f is adaptive submodular if the expected marginal benefits $\Delta(t \mid \mathbf{x}_\mathcal{A})$ of adding a new test t can only decrease as we gather more observations. f is called *strongly adaptively monotone* w.r.t. P if, informally, "observations never hurt" with respect to the expected reward. Formally, for all $\mathcal{A}$, all $t \notin \mathcal{A}$, and all $x \in \mathcal{X}$ we require $\mathbb{E}_H[f(\mathcal{A}, H) \mid \mathbf{x}_\mathcal{A}] \leq \mathbb{E}_H[f(\mathcal{A} \cup \{t\}, H) \mid \mathbf{x}_\mathcal{A}, X_t = x]$.

The performance guarantee for GBS follows from the following general result about the greedy algorithm for adaptive submodular functions (applied with $Q = 1$ and $\eta = p_{\min}$):

Theorem 1 (Theorem 10 of [9] with $\alpha = 1$). *Suppose $f : 2^\mathcal{T} \times \mathcal{H} \to \mathbb{R}_{\geq 0}$ is adaptive submodular and strongly adaptively monotone with respect to P and there exists Q such that $f(\mathcal{T}, h) = Q$ for all h. Let η be any value such that $f(S, h) > Q - \eta$ implies $f(S, h) = Q$ for all sets S and hypotheses h. Then for self–certifying instances the adaptive greedy policy π satisfies $c(\pi) \leq c(\pi^*)\left(\ln\left(\frac{Q}{\eta}\right) + 1\right)$.*

The technical requirement that instances be *self–certifying* means that the policy will have proof that it has obtained the maximum possible objective value, Q, immediately upon doing so. It is not difficult to show that this is the case with the instances we consider in this paper. We refer the interested reader to [9] for more detail.

In the following sections, we will use the concept of adaptive submodularity to provide the first approximation guarantees for Bayesian active learning with noisy observations.

3 The Equivalence Class Determination Problem and the EC^2 Algorithm

We now wish to consider the Bayesian active learning problem where tests can have noisy outcomes. Our general strategy is to reduce the problem of noisy observations to the noiseless setting. To gain intuition, consider a simple model where tests have binary outcomes, and we know that the outcome of exactly one test, chosen uniformly at random unbeknown to us, is flipped. If any pair of hypotheses $h \neq h'$ differs by the outcome of at least three tests, we can still uniquely determine the correct hypothesis after running all tests. In this case we can reduce the noisy active learning problem to the noiseless setting by, for each hypothesis, creating N "noisy" copies, each obtained by flipping the outcome of one of the N tests. The modified prior P' would then assign mass $P'(h') = P(h)/N$ to each noisy copy h' of h. The conditional distribution $P'(\mathbf{X}_\mathcal{T} \mid h')$ is still deterministic (obtained by flipping the outcome of one of the tests). Thus, each hypothesis h_i in the original problem is now associated with a set $\mathcal{H}_i$ of hypotheses in the modified problem instance. However, instead of selecting tests to determine which noisy copy has been realized, we only care which set $\mathcal{H}_i$ is realized.

3

The Equivalence Class Determination problem (ECD). More generally, we introduce the *Equivalence Class Determination problem*[4], where our set of hypotheses $\mathcal{H}$ is partitioned into a set of m equivalence classes $\{\mathcal{H}_1, \ldots, \mathcal{H}_m\}$ so that $\mathcal{H} = \biguplus_{i=1}^{m} \mathcal{H}_i$, and the goal is to determine which class $\mathcal{H}_i$ the true hypothesis lies in. Formally, upon termination with observations $\mathbf{x}_\mathcal{A}$ we require that $\mathcal{V}(\mathbf{x}_\mathcal{A}) \subseteq \mathcal{H}_i$ for some i. As with the ODT problem, the goal is to minimize the expected cost of the tests, where the expectation is taken over the true hypothesis sampled from P. In §4, we will show how the Equivalence Class Determination problem arises naturally from Bayesian experimental design problems in probabilistic models.

Given the fact that GBS performs near-optimally on the Optimal Decision Tree problem, a natural approach to solving ECD would be to run GBS until the termination condition is met. Unfortunately, and perhaps surprisingly, GBS can perform very poorly on the ECD problem. Consider an instance with a uniform prior over n hypotheses, $h_1, \ldots, h_n$, and two equivalence classes $\mathcal{H}_1 := \{h_i : 1 \leq i < n\}$ and $\mathcal{H}_2 := \{h_n\}$. There are tests $\mathcal{T} = \{1, \ldots, n\}$ such that $h_i(t) = \mathbf{1}[i = t]$, all of unit cost. Hereby, $\mathbf{1}[\Lambda]$ is the indicator variable for event Λ. In this case, the optimal policy only needs to select test n, however GBS may select tests $1, 2, \ldots, n$ in order until running test t, where $H = h_t$ is the true hypothesis. Given our uniform prior, it takes $n/2$ tests in expectation until this happens, so that GBS pays, in expectation, $n/2$ times the optimal expected cost in this instance.

The poor performance of GBS in this instance may be attributed to its lack of consideration for the equivalence classes. Another natural heuristic would be to run the greedy information gain policy, only with the entropy measured with respect to the probability distribution on *equivalence classes* rather than hypotheses. Call this policy π_{IG}. It is clearly aware of the equivalence classes, as it adaptively and myopically selects tests to reduce the uncertainty of the realized class, measured w.r.t. the Shannon entropy. However, we can show there are instances in which it pays $\Omega(n/\log(n))$ times the optimal cost, even under a uniform prior. Refer to Appendix B for details.

The EC^2 algorithm. The reason why GBS fails is because reducing the version space mass does not necessarily facilitate differentiation among the classes $\mathcal{H}_i$. The reason why π_{IG} fails is that there are complementarities among tests; a set of tests can be far better than the sum of its parts. Thus, we would like to optimize an objective function that encourages differentiation among classes, but lacks complementarities. We adopt a very elegant idea from Dasgupta [8], and define weighted edges between hypotheses that we aim to distinguish between. However, instead of introducing edges between arbitrary pairs of hypotheses (as done in [8]), we only introduce edges between hypotheses in different classes. Tests will allow us to cut edges inconsistent with their outcomes, and we aim to eliminate all inconsistent edges while minimizing the expected cost incurred. We now formalize this intuition.

Specifically, we define a set of edges $\mathcal{E} = \cup_{1 \leq i < j \leq m} \{\{h, h'\} : h \in \mathcal{H}_i, h' \in \mathcal{H}_j\}$, consisting of all (unordered) pairs of hypotheses belonging to distinct classes. These are the edges that must be *cut*, by which we mean for any edge $\{h, h'\} \in \mathcal{E}$, at least one hypothesis in $\{h, h'\}$ must be ruled out (i.e., eliminated from the version space). Hence, a test t run under true hypothesis h is said to cut edges $\mathcal{E}_t(h) := \{\{h', h''\} : h'(t) \neq h(t) \text{ or } h''(t) \neq h(t)\}$. See Fig. 1(a) for an illustration. We define a weight function $w : \mathcal{E} \to \mathbb{R}_{\geq 0}$ by $w(\{h, h'\}) := P(h) \cdot P(h')$. We extend the weight function to an additive (modular) function on sets of edges in the natural manner, i.e., $w(\mathcal{E}') := \sum_{e \in \mathcal{E}'} w(e)$. The objective f_{EC} that we will greedily maximize is then defined as the weight of the edges cut (EC):

$$f_{EC}(\mathcal{A}, h) := w\Big(\bigcup_{t \in \mathcal{A}} \mathcal{E}_t(h)\Big) \tag{3.1}$$

The key insight that allows us to prove approximation guarantees for f_{EC} is that f_{EC} shares the same beneficial properties that make f_{GBS} amenable to efficient greedy optimization. We prove this fact, as stated in Proposition 2, in Appendix A.

Proposition 2. *The objective f_{EC} is strongly adaptively monotone and adaptively submodular.*

Based on the objective f_{EC}, we can calculate the marginal benefits for test t upon observations $\mathbf{x}_\mathcal{A}$ as

$$\Delta_{\mathrm{EC}}(t | \mathbf{x}_\mathcal{A}) := \mathbb{E}_H[f_{EC}(\mathcal{A} \cup \{t\}, H) - f_{EC}(\mathcal{A}, H) \mid \mathbf{x}_\mathcal{A}].$$

---

[4]Bellala et al. simultaneously studied ECD [2], and, like us, used it to model active learning with noise [3]. They developed an extension of GBS for ECD. We defer a detailed comparison of our approaches to future work.

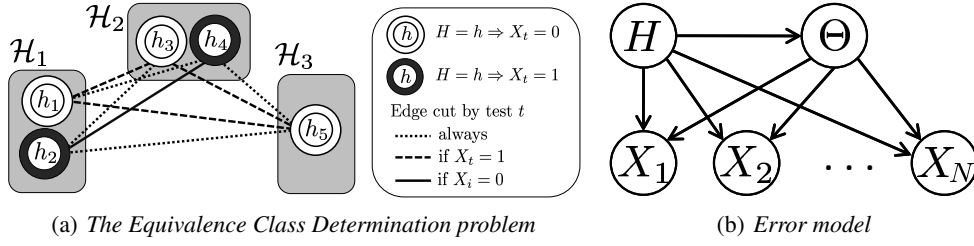(a) *The Equivalence Class Determination problem*   (b) *Error model*

Figure 1: (a) An instance of Equivalence Class Determination with binary test outcomes, shown with the set of edges that must be cut, and depicting the effects of test i under different outcomes. (b) The graphical model underlying our error model.

We call the adaptive policy π_{EC} that, after observing $\mathbf{x}_{\mathcal{A}}$, greedily selects test $t^* \in \arg\max_t \Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})\,/c(t)$, the EC^2 algorithm (for *equivalence class edge cutting*).

Note that these instances are self–certifying, because we obtain maximum objective value if and only if the version space lies within an equivalence class, and the policy can certify this condition when it holds. So we can apply Theorem 1 to show EC^2 obtains a $\ln(Q/\eta) + 1$ approximation to Equivalence Class Determination. Hereby, $Q = w(\mathcal{E}) = 1 - \sum_i (P(h \in \mathcal{H}_i))^2 \leq 1$ is the total weight of all edges that need to be cut, and $\eta = \min_{e \in \mathcal{E}} w(e) \geq p_{\min}^2$ is a bound on the minimum weight among all edges. We have the following result:

Theorem 3. *Suppose $P(h)$ is rational for all $h \in \mathcal{H}$. For the adaptive greedy policy π_{EC} implemented by EC^2 it holds that*

$$c(\pi_{EC}) \leq (2\ln(1/p_{\min}) + 1)c(\pi^*),$$

where $p_{\min} := \min_{h \in \mathcal{H}} P(h)$ is the minimum prior probability of any hypothesis, and π^* is the optimal policy for the Equivalence Class Determination problem.

In the case of unit cost tests, we can apply a technique of Kosaraju et al. [14], originally developed for the GBS algorithm, to improve the approximation guarantee to $\mathcal{O}(\log n)$ by applying EC^2 with a modified prior distribution. We defer details to the full version of this paper.

A Fast Implementation of EC^2. The time running time of EC^2 is dominated by the evaluations of $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$. The naive way to compute $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$ is to construct a graph on the n hypotheses with weighted edges as prescribed by EC^2, and then see which edges are cut by t for each potential test outcome. Assuming there are ℓ possible outcomes of a test, and that we can evaluate $h(t)$ in unit time for all h and t, this will take $\mathcal{O}(n^2\ell)$ time. With N tests, the total time per round of EC^2 is $\mathcal{O}(Nn^2\ell)$. However, there is a much faster way to compute $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$. Note that $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$ equals

$$\mathbb{E}_{x_t \sim X_t|\mathbf{x}_{\mathcal{A}}} \left[\frac{1}{2} \sum_{i \neq j} \Big(P\left(\mathcal{H}_i \cap \mathcal{V}(\mathbf{x}_{\mathcal{A}})\right) P\left(\mathcal{H}_j \cap \mathcal{V}(\mathbf{x}_{\mathcal{A}})\right) - P\left(\mathcal{H}_i \cap \mathcal{V}(\mathbf{x}_{\mathcal{A}}, x_t)\right) P\left(\mathcal{H}_j \cap \mathcal{V}(\mathbf{x}_{\mathcal{A}}, x_t)\right) \Big) \right].$$

Now, compute $\alpha(i, x_t) := P\left(\mathcal{H}_i \cap \mathcal{V}(\mathbf{x}_{\mathcal{A}}, x_t)\right)$ for each i and x, then compute $\beta(i) := P\left(\mathcal{H}_i\right) = \sum_x \alpha(i, x)$. Next, compute $\gamma(x_t) := P\left(x_t \mid \mathbf{x}_{\mathcal{A}}\right) = \sum_i \alpha(i, x_t)/ \sum_i \beta(i)$. All of these terms can be computed in total time $\mathcal{O}(n)$ by iterating over the hypotheses and for each h adding $P(h)$ to the appropriate terms (i.e., $\beta(i)$, $\alpha(i, x)$, and $\gamma(x)$ if $h \in \mathcal{H}_i$ and $h(t) = x$). Using these variables, we can rewrite $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$ as $\mathbb{E}_{x_t \sim X_t|\mathbf{x}_{\mathcal{A}}} \left[\frac{1}{2} \sum_{i \neq j} \Big(\beta(i)\beta(j) - \alpha(i, x_t)\alpha(j, x_t) \Big) \right]$. Note that for any $\eta_1, \eta_2, \ldots, \eta_m \in \mathbb{R}$, we have $\sum_{i \neq j} \eta_i \eta_j = \left(\sum_i \eta_i \right)^2 - \sum_i \eta_i^2$. Using this equality, we can evaluate sums like $\sum_{i \neq j} \Big(\beta(i)\beta(j) - \alpha(i, x_t)\alpha(j, x_t) \Big)$ in $\mathcal{O}(m)$ time, where there are m equivalence classes. Hence the total time to evaluate $\Delta_{\text{EC}}(t\,|\,\mathbf{x}_{\mathcal{A}})$ is $\mathcal{O}(n + m\ell)$ using this method. In a similar manner, we can reduce the running time still further to $\mathcal{O}(n)$, by incrementally computing terms such as $\left(\sum_i \alpha(i, x) \right)^2$ and $\sum_i \alpha(i, x)^2$ as we iterate through the hypotheses. The total time per round of EC^2 is then $\mathcal{O}(Nn)$. Additionally, the number of evaluations the algorithm needs to make can often be significantly reduced in practice using the *accelerated adaptive greedy* algorithm, as discussed in [9].

4   Bayesian Active Learning with Noise and the EFFECXTIVE Algorithm

We now address the case of noisy observations, using ideas from §3. With noisy observations, the conditional distribution $P(X_1, \ldots, X_N \mid h)$ is no longer deterministic. We model the noise using an additional random variable Θ. Fig. 1(b) depicts the underlying graphical model. The vector of test outcomes $\mathbf{x}_{\mathcal{T}}$ is assumed to be an arbitrary, deterministic function $\mathbf{x}_{\mathcal{T}} : \mathcal{H} \times \mathrm{supp}(\Theta) \to \mathcal{X}^N$; hence $\mathbf{X}_{\mathcal{T}} \mid h$ is distributed as $\mathbf{x}_{\mathcal{T}}(h, \Theta_h)$ where Θ_h is distributed as $P(\theta \mid h)$. For example, there might be up to $s = |\mathrm{supp}(\Theta)|$ ways any particular disease could manifest itself, with different patients with the same disease suffering from different symptoms.

In cases where it is always possible to identify the true hypothesis, i.e., $\mathbf{x}_{\mathcal{T}}(h, \theta) \neq \mathbf{x}_{\mathcal{T}}(h', \theta')$ for all $h \neq h'$ and all $\theta, \theta' \in \mathrm{supp}(\Theta)$, we can reduce the problem to Equivalence Class Determination with hypotheses $\{\mathbf{x}_{\mathcal{T}}(h, \theta) : h \in \mathcal{H}, \theta \in \mathrm{supp}(\Theta)\}$ and equivalence classes $\mathcal{H}_i := \{\mathbf{x}_{\mathcal{T}}(h_i, \theta) : \theta \in \mathrm{supp}(\Theta)\}$ for all i. Then Theorem 3 immediately yields that the approximation factor of EC^2 is at most $2 \ln \left(1/ \min_{h,\theta} P(h, \theta) \right) + 1$, where the minimum is taken over all (h, θ) in the support of P. In the unit cost case, running EC^2 with a modified prior à la Kosaraju et al. [14] allows us to obtain an $\mathcal{O}(\log |\mathcal{H}| + \log |\mathrm{supp}(\Theta)|)$ approximation factor. Note this model allows us to incorporate noise with complex correlations.

However, a major challenge when dealing with noisy observations is that it is not always possible to distinguish distinct hypotheses. Even after we have run all tests, there will generally still be uncertainty about the true hypothesis, i.e., the posterior distribution $P(H \mid \mathbf{x}_{\mathcal{T}})$ obtained using Bayes' rule may still assign non-zero probability to more than one hypothesis. If so, uniquely determining the true hypothesis is not possible. Instead, we imagine that there is a set $\mathcal{D}$ of possible *decisions* we may make after (adaptively) selecting a set of tests to perform and we must choose one (e.g., we must decide how to treat the medical patient, which scientific theory to adopt, or which classifier to use, given our observations). Thus our goal is to gather data to make effective decisions [12]. Formally, for any decision $d \in \mathcal{D}$ we take, and each realized hypothesis h, we incur some loss $\ell(d, h)$. Decision theory recommends, after observing $\mathbf{x}_{\mathcal{A}}$, to choose the decision d^* that minimizes the *risk*, i.e., the expected loss, namely $d^* \in \arg \min_d \mathbb{E}_H[\ell(d, H) \mid \mathbf{x}_{\mathcal{A}}]$.

A natural goal in Bayesian active learning is thus to adaptively pick observations, until we are guaranteed to make the same decision (and thus incur the same expected loss) that we would have made had we run *all* tests. Thus, we can reduce the noisy Bayesian active learning problem to the ECD problem by defining the equivalence classes over all test outcomes that lead to the same minimum risk decision. Hence, for each decision $d \in \mathcal{D}$, we define

$$\mathcal{H}_d := \{\mathbf{x}_{\mathcal{T}} : d = \arg \min_{d'} \mathbb{E}_H[\ell(d', H) \mid \mathbf{x}_{\mathcal{T}}]\}. \tag{4.1}$$

If multiple decisions minimize the risk for a particular $\mathbf{x}_{\mathcal{T}}$, we break ties arbitrarily. Identifying the best decision $d \in \mathcal{D}$ then amounts to identifying which equivalence class $\mathcal{H}_d$ contains the realized vector of outcomes, which is an instance of ECD.

One common approach to this problem is to myopically pick tests maximizing the decision-theoretic *value of information* (VoI): $\Delta_{\mathrm{VoI}}(t \mid \mathbf{x}_{\mathcal{A}}) := \min_d \mathbb{E}_H[\ell(d, H) \mid \mathbf{x}_{\mathcal{A}}] - \mathbb{E}_{x_t \sim X_t \mid \mathbf{x}_{\mathcal{A}}}[\min_d \mathbb{E}_H[\ell(d, H) \mid \mathbf{x}_{\mathcal{A}}, x_t]]$. The VoI of a test t is the expected reduction in the expected loss of the best decision due to the observation of x_t. However, we can show there are instances in which such a policy pays $\Omega(n / \log(n))$ times the optimal cost, even under a uniform prior on (h, θ) and with $|\mathrm{supp}(\Theta)| = 2$. Refer to Appendix B for details. In contrast, on such instances EC^2 obtains an $\mathcal{O}(\log n)$ approximation. More generally, we have the following result for EC^2 as an immediate consequence of Theorem 3.

Theorem 4. *Fix hypotheses $\mathcal{H}$, tests $\mathcal{T}$ with costs $c(t)$ and outcomes in $\mathcal{X}$, decision set $\mathcal{D}$, and loss function ℓ. Fix a prior $P(H, \Theta)$ and a function $\mathbf{x}_{\mathcal{T}} : \mathcal{H} \times \mathrm{supp}(\Theta) \to \mathcal{X}^N$ which define the probabilistic noise model. Let $c(\pi)$ denote the expected cost of π incurs to find the best decision, i.e., to identify which equivalence class $\mathcal{H}_d$ the outcome vector $\mathbf{x}_{\mathcal{T}}$ belongs to. Let π^* denote the policy minimizing $c(\cdot)$, and let π_{EC} denote the adaptive policy implemented by EC^2. Then it holds that*

$$c(\pi_{EC}) \leq \left(2 \ln \left(\frac{1}{p'_{\min}} \right) + 1 \right) c(\pi^*),$$

where $p'_{\min} := \min_{h \in \mathcal{H}} \{P(h, \theta) : P(h, \theta) > 0\}$.

6

If all tests have unit cost, by using a modified prior [14] the approximation factor can be improved to $\mathcal{O}\left(\log|\mathcal{H}| + \log|\operatorname{supp}(\Theta)|\right)$ as in the case of Theorem 3.

The EFFECXTIVE **algorithm.** For some noise models, Θ may have exponentially–large support. In this case reducing Bayesian active learning with noise to Equivalence Class Determination results in instances with exponentially-large equivalence classes. This makes running EC^2 on them challenging, since explicitly keeping track of the equivalence classes is impractical. To overcome this challenge, we develop EFFECXTIVE, a particularly efficient algorithm which approximates EC^2.

For clarity, we only consider the $0-1$ loss, i.e., our goal is to find the most likely hypothesis (MAP estimate) given all the data $\mathbf{x}_{\mathcal{T}}$, namely $h^*(\mathbf{x}_{\mathcal{T}}) := \arg\max_h P(h \mid \mathbf{x}_{\mathcal{T}})$. Recall definition (4.1), and consider the weight of edges between distinct equivalence classes $\mathcal{H}_i$ and $\mathcal{H}_j$:

$$w(\mathcal{H}_i \times \mathcal{H}_j) = \sum_{\mathbf{x}_{\mathcal{T}} \in \mathcal{H}_i, \mathbf{x}'_{\mathcal{T}} \in \mathcal{H}_j} P(\mathbf{x}_{\mathcal{T}})P(\mathbf{x}'_{\mathcal{T}}) = \Big(\sum_{\mathbf{x}_{\mathcal{T}} \in \mathcal{H}_i} P(\mathbf{x}_{\mathcal{T}})\Big)\Big(\sum_{\mathbf{x}'_{\mathcal{T}} \in \mathcal{H}_j} P(\mathbf{x}'_{\mathcal{T}})\Big) = P(\mathbf{X}_{\mathcal{T}} \in \mathcal{H}_i)P(\mathbf{X}_{\mathcal{T}} \in \mathcal{H}_j).$$

In general, $P(\mathbf{X}_{\mathcal{T}} \in \mathcal{H}_i)$ can be estimated to arbitrary accuracy using a rejection sampling approach with bounded sample complexity. We defer details to a longer version of the paper. Here, we focus on the case where, upon observing all tests, the hypothesis is uniquely determined, i.e., $P(H \mid \mathbf{x}_{\mathcal{T}})$ is deterministic for all $\mathbf{x}_{\mathcal{T}}$ in the support of P. In this case, it holds that $P(\mathbf{X}_{\mathcal{T}} \in \mathcal{H}_i) = P(H = h_i)$. Thus, the total weight is

$$\sum_{i \neq j} w(\mathcal{H}_i \times \mathcal{H}_j) = \Big(\sum_i P(h_i)\Big)^2 - \sum_i P(h_i)^2 = 1 - \sum_i P(h_i)^2.$$

This insight motivates us to use the objective function

$$\Delta_{\text{Eff}}(t\,|\,\mathbf{x}_{\mathcal{A}}) := \Big[\sum_x P(X_t = x \mid \mathbf{x}_{\mathcal{A}})\Big(\sum_i P(h_i \mid \mathbf{x}_{\mathcal{A}}, X_t = x)^2\Big)\Big] - \sum_i P(h_i \mid \mathbf{x}_{\mathcal{A}})^2,$$

which is the expected reduction in weight from the prior to the posterior distribution. Note that the weight of a distribution $1 - \sum_i P(h_i)^2$ is a monotonically increasing function of the Rényi entropy (of order 2), which is $-\frac{1}{2}\log\sum_i P(h_i)^2$. Thus the objective Δ_{Eff} can be interpreted as a (non-standard) information gain in terms of the (exponentiated) Rényi entropy. In our experiments, we show that this criterion performs well in comparison to existing experimental design criteria, including the classical Shannon information gain. Computing $\Delta_{\text{Eff}}(t\,|\,\mathbf{x}_{\mathcal{A}})$ requires us to perform one inference task for each outcome x of X_t, and $\mathcal{O}(n)$ computations to calculate the weight for each outcome. We call the algorithm that greedily optimizes Δ_{Eff} the EFFECXTIVE algorithm (since it uses an Efficient Edge Cutting approXimate objective), and present pseudocode in Algorithm 1.

Input: Set of hypotheses $\mathcal{H}$; Set of tests $\mathcal{T}$; prior distribution P; function f.
begin
    $\mathcal{A} \leftarrow \emptyset$;
    while $\exists h \neq h' : P(h \mid \mathbf{x}_{\mathcal{A}}) > 0$ *and* $P(h' \mid \mathbf{x}_{\mathcal{A}}) > 0$ **do**
        foreach $t \in \mathcal{T} \setminus \mathcal{A}$ **do**
            $\Delta_{\text{Eff}}(t\,|\,\mathbf{x}_{\mathcal{A}}) := \Big[\sum_x P(X_t = x \mid \mathbf{x}_{\mathcal{A}})\big(\sum_i P(h_i \mid \mathbf{x}_{\mathcal{A}}, X_t = x)^2\big)\Big] - \sum_i P(h_i \mid \mathbf{x}_{\mathcal{A}})^2$;
        Select $t^* \in \arg\max_t \Delta_{\text{Eff}}(t\,|\,\mathbf{x}_{\mathcal{A}})\,/c(t)$; Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{t^*\}$ and observe outcome x_{t^*};
end

Algorithm 1: The EFFECXTIVE algorithm using the Efficient Edge Cutting approXimate objective.

---

5 Experiments

Several economic theories make claims to explain how people make decisions when the payoffs are uncertain. Here we use human subject experiments to compare four key theories proposed in literature. The uncertainty of the payoff in a given situation is represented by a *lottery* L, which is simply a random variable with a range of *payoffs* $\mathcal{L} := \{\ell_1, \ldots, \ell_k\}$. For our purposes, a payoff is an integer denoting how many dollars you receive (or lose, if the payoff is negative). Fix lottery L, and let $p_i := \mathbb{P}[L = \ell_i]$. The four theories posit distinct utility functions, with agents preferring larger utility lotteries. Three of the theories have associated parameters. The *Expected Value*

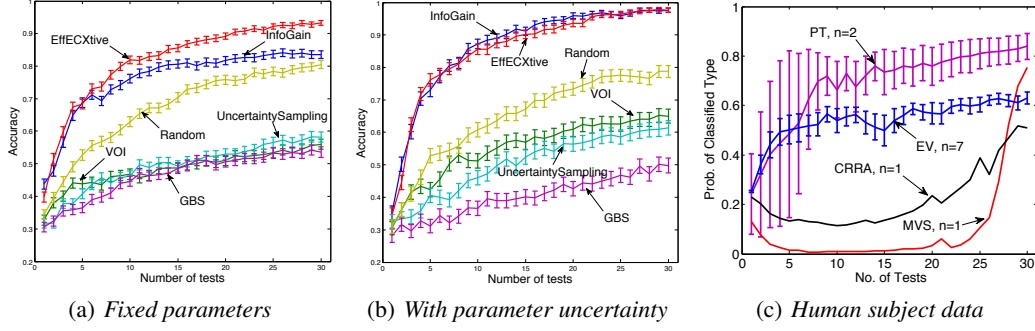|(a) *Fixed parameters*|(b) *With parameter uncertainty*|(c) *Human subject data*|

Figure 2: (a) Accuracy of identifying the true model with fixed parameters, (b) Accuracy using a grid of parameters, incorporating uncertainty in their values, (c) Experimental results: 11 subjects were classified into the theories that described their behavior best. We plot probability of classified type.

theory [21] posits simply $U_{EV}(L) = \mathbb{E}[L]$, and has no parameters. *Prospect theory* [13] posits $U_{PT}(L) = \sum_i f(\ell_i)w(p_i)$ for nonlinear functions $f(\ell_i) = \ell_i^\rho$, if $\ell_i \geq 0$ and $f(\ell_i) = -\lambda(-\ell_i)^\rho$, if $\ell_i < 0$, and $w(p_i) = e^{-(\log(1/p_i))^\alpha}$ [20]. The parameters $\Theta_{PT} = \{\rho, \lambda, \alpha\}$ represent risk aversion, loss aversion and probability weighing factor respectively. For Portfolio optimization problems, financial economists have used value functions that give weights to different moments of the lottery [11]: $U_{MVS}(L) = w_\mu\mu - w_\sigma\sigma + w_\nu\nu$, where $\Theta_{MVS} = \{w_\mu, w_\sigma, w_\nu\}$ are the weights for the mean, standard deviation and standardized skewness of the lottery respectively. In *Constant Relative Risk Aversion theory* [19], there is a parameter $\Theta_{CRRA} = a$ representing the level of risk aversion, and the utility posited is $U_{CRRA}(L) = \sum_i p_i\ell_i^{1-a}/(1-a)$ if $a \neq 1$, and $U_{CRRA}(L) = \sum_i p_i\log(\ell_i)$, if $a = 1$.

The goal is to adaptively select a sequence of tests to present to a human subject in order to distinguish which of the four theories best explains the subject's responses. Here a test t is a pair of lotteries, (L_1^t, L_2^t). Based on the theory that represents behaviour, one of the lotteries would be preferred to the other, denoted by a binary response $x_t \in \{1, 2\}$. The possible payoffs were fixed to $\mathcal{L} = \{-10, 0, 10\}$ (in dollars), and the distribution (p_1, p_2, p_3) over the payoffs was varied, where $p_i \in \{0.01, 0.99\} \cup \{0.1, 0.2, \ldots, 0.9\}$. By considering all non-identical pairs of such lotteries, we obtained the set of possible tests.

We compare five different methods: EFFECXTIVE, greedily maximizing Information Gain (IG), Uncertainty Sampling[5] (US), minimizing Version Space (VS), and tests selected at Random. We first evaluated the ability of the algorithms to recover the true model based on simulated responses. We chose parameter values for the theories, such that they made distinct predictions, and were consistent with the values proposed in literature [13]. We drew 1000 samples of the true model and fixed the parameters of the model to some canonical values, $\Theta_{PT} = \{0.9, 2.2, 0.9\}, \Theta_{MVS} = \{0.8, 0.25, 0.25\}, \Theta_{CRRA} = 1$. Responses were generated using a softmax function, with the probability of response $x_t = 1$ given by $P(x_t = 1) = 1/(1 + e^{U(L_2^t)-U(L_1^t)})$. Fig. 2(a) shows the performance of the 5 methods, in terms of the accuracy of recovering the true model with the number of tests. We find that US and VS both perform significantly worse than Random in the presence of noise. EFFECXTIVE outperforms InfoGain significantly, which outperforms Random.

We also considered uncertainty in the values of the parameters, by setting ρ from 0.85-0.95, λ from 2.1-2.3, α from 0.9-1; w_μ from 0.8-1.0, w_σ from 0.2-0.3, w_ν from 0.2-0.3; and a from 0.9-1.0, all with 3 values per parameter. We generated 500 random samples by first randomly sampling a model and then randomly sampling parameter values. EFFECXTIVE and InfoGain outperformed Random significantly, Fig. 2(b), although InfoGain did marginally better among the two. The increased parameter range potentially poses model identifiability issues, and violates some of the assumptions behind EFFECXTIVE, decreasing its performance to the level of InfoGain.

After obtaining informed consent, we tested 11 human subjects to determine which model fit their behaviour best. Laboratory experiments have been used previously to distinguish economic theories, [4], and here we used a real-time, dynamically optimized experiment that required fewer tests. Subjects were presented 30 tests using EFFECXTIVE. To incentivise the subjects, one of these tests

---
[5]Uncertainty sampling greedily selects the test whose outcome distribution has maximum Shannon entropy.

was picked at random, and subjects received payment based the outcome of their chosen lottery. The behavior of most subjects (7 out of 10) was best described by EV. This is not unexpected given the unusually high quantitative abilities of the subjects. We also found heterogeneity in classification: One subject got classified as MVS, as identified by her violations of stochastic dominance in the last few tests. 2 subjects were best described by prospect theory since they exhibited a high degree of loss aversion and risk aversion. One subject was also classified as a CRRA-type (log-utility maximizer). Figure 2(c) shows the probability of the classified model with number of tests. Although we need a bigger and more diverse sample to make significant claims of the validity of different economic theories, our preliminary results indicate that subject types can be identified and there is heterogeneity in the population. It also serves as a practical application of real-time dynamic experimental design optimization that is necessary to collect data on human economic behavior.

6 Conclusions

In this paper, we considered the problem of adaptively selecting which noisy tests to perform in order to identify an unknown hypothesis sampled from a known prior distribution. We studied the Equivalence Class Determination problem as a means to reduce the case of noisy observations to the classic, noiseless case. We introduced EC^2, an adaptive greedy algorithm that is guaranteed to choose the same hypothesis as if it had observed the outcome of all tests, and incurs near-minimal expected cost among all policies with this guarantee. This is in contrast to popular heuristics that are greedy w.r.t. version space mass reduction, information gain or value of information, all of which we show can be very far from optimal. EC^2 works by greedily optimizing an objective tailored to differentiate between sets of observations that lead to different decisions. Our bounds rely on the fact that this objective function is adaptive submodular. We also develop EFFECXTIVE, a practical algorithm based on EC^2, that can be applied to arbitrary probabilistic models in which efficient exact inference is possible. We apply EFFECXTIVE to a Bayesian experimental design problem, and our results indicate its effectiveness in comparison to existing algorithms. We believe that our results provide an interesting direction towards providing a theoretical foundation for practical active learning and experimental design problems.

References

[1] N. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

[2] Gowtham Bellala, Suresh Bhavnani, and Clayton Scott. Extensions of generalized binary search to group identification and exponential costs. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[3] Gowtham Bellala, Suresh K. Bhavnani, and Clayton D. Scott. Group-based query learning for rapid diagnosis in time-critical situations. *CoRR*, abs/0911.4511, 2009.

[4] Colin F. Camerer. An experimental test of several generalized utility theories. *The Journal of Risk and Uncertainty*, 2(1):61–104, 1989.

[5] V. T. Chakaravarthy, V. Pandit, S. Roy, P. Awasthi, and M. Mohania. Decision trees for entity identification: Approximation algorithms and hardness results. In *In Proceedings of the ACM- SIGMOD Symposium on Principles of Database Systems*, 2007.

[6] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, Aug. 1995.

[7] Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, 2004.

[8] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 235–242. MIT Press, Cambridge, MA, 2006.

[9] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *CoRR*, abs/1003.3967v3, 2010.

[10] Andrew Guillory and Jeff Bilmes. Average-case active learning with costs. In *The 20th International Conference on Algorithmic Learning Theory*, University of Porto, Portugal, October 2009.

[11] Giora Hanoch and Haim Levy. Efficient portfolio selection with quadratic and cubic utility. *The Journal of Business*, 43(2):181–189, 1970.

[12] R. A. Howard. Information value theory. In *IEEE Transactions on Systems Science and Cybernetics (SSC-2)*, 1966.

[13] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[14] S. Rao Kosaraju, Teresa M. Przytycka, and Ryan S. Borgstrom. On an optimal split tree problem. In *WADS '99: Proceedings of the 6th International Workshop on Algorithms and Data Structures*, pages 157–168, London, UK, 1999. Springer-Verlag.

[15] Andreas Krause and Carlos Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research (JAIR)*, 35:557–591, 2009.

[16] D. V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

[17] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[18] Rob Nowak. Noisy generalized binary search. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1366–1374. 2009.

[19] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1):122–136, 1964.

[20] D. Prelec. The probablity weighting function. *Econometrica*, 66(3):497–527, 1998.

[21] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1947.

[22] Alice X. Zheng, Irina Rish, and Alina Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, 2005.

A   Additional Proofs

Lemma 5. *The objective function f of Eq. (3.1) is strongly adaptive monotone.*

Proof. We must show that for all $\mathbf{x}_{\mathcal{A}}$, $t \notin \mathcal{A}$ and possible answer x for test t that

$$\mathbb{E}_H[f(\mathcal{A}, H) \mid \mathbf{x}_{\mathcal{A}}] \leq \mathbb{E}_H[f(\mathcal{A} \cup \{t\}, H) \mid \mathbf{x}_{\mathcal{A}}, X_t = x] \tag{A.1}$$

Towards this end, it is useful to notice that for all $t \in \mathcal{T}$ the function $h \mapsto \mathcal{E}_t(h)$ depends only on X_t. Hence for any $\mathbf{x}_{\mathcal{A}}$, the function $h \mapsto f(\mathcal{A}, h)$ is constant over realizations $\mathbf{x}_{\mathcal{T}} \succ \mathbf{x}_{\mathcal{A}}$, so we can define a function $g(\mathbf{x}_{\mathcal{A}})$ such that $g(\mathbf{x}_{\mathcal{A}}) = \mathbb{E}_H[f(\mathcal{A}, H) \mid \mathbf{x}_{\mathcal{A}}]$ by $g(\mathbf{x}_{\mathcal{A}}) := w\left(\bigcup_{t \in \mathcal{A}} \mathcal{E}_t(x_t)\right)$ where $\mathbf{x}_{\mathcal{A}} = (x_t)_{t \in \mathcal{A}}$ and $\mathcal{E}_t(x)$ is the set of edges cut by t if $X_t = x$. Note that for all $\mathbf{x}_{\mathcal{A}} \prec \mathbf{x}_{\mathcal{B}}$ we have $g(\mathbf{x}_{\mathcal{A}}) \leq g(\mathbf{x}_{\mathcal{B}})$, since the edge weights are nonnegative. Setting $\mathcal{B} = \mathcal{A} \cup \{t\}$ yields Eq. (A.1) and hence implies strong adaptive monotonicity. $\square$

Lemma 6. *The objective function f of Eq. (3.1) is adaptive submodular for any prior with rational values.*

Proof. We first prove the result assuming a uniform prior $P(\cdot)$, and then show how to reduce the general prior case to the uniform prior case. Hence all edges have weight $1/n^2$, where there are n hypotheses. For convenience, we also rescale our units of reward so that all edges have unit weight. (Note that f is adaptive submodular iff cf is for any constant $c > 0$.) To prove adaptive submodularity, we must show that for all $\mathbf{x}_{\mathcal{A}} \prec \mathbf{x}_{\mathcal{B}}$ and $t \in \mathcal{T}$, we have $\Delta(t \mid \mathbf{x}_{\mathcal{B}}) \leq \Delta(t \mid \mathbf{x}_{\mathcal{A}})$. Fix t and $\mathbf{x}_{\mathcal{A}}$, and let $\mathcal{V}(\mathbf{x}_{\mathcal{A}}) := \{h : P(h \mid \mathbf{x}_{\mathcal{A}}) > 0\}$ denote the version space, if $\mathbf{x}_{\mathcal{A}}$ encodes the observed outcomes. Let $n_{\mathcal{V}} := |\mathcal{V}(\mathbf{x}_{\mathcal{A}})|$ be the number of hypotheses in the version space. Likewise, let $n_{i,a}(\mathbf{x}_{\mathcal{A}}) := |\{h : h \in \mathcal{V}(\mathbf{x}_{\mathcal{A}}, X_t = a) \cap \mathcal{H}_i\}|$, and let $n_a(\mathbf{x}_{\mathcal{A}}) := \sum_{i=1}^{m} n_{i,a}(\mathbf{x}_{\mathcal{A}})$. Also, define $e_a(\mathbf{x}_{\mathcal{A}}) := \frac{1}{2} \sum_{i \neq j} \sum_{b \neq a} n_{i,b}(\mathbf{x}_{\mathcal{A}}) \cdot n_{j,b}(\mathbf{x}_{\mathcal{A}})$ to be the number of edges cut such that at t both hypotheses agree with each other but disagree with the realized hypothesis h^*, conditioning on $X_t = a$. We define a function θ of these quantities such that $\Delta(t \mid \mathbf{x}_{\mathcal{A}}) = \theta(\mathbf{n}(\mathbf{x}_{\mathcal{A}}), \mathbf{e}(\mathbf{x}_{\mathcal{A}}))$, where $\mathbf{n}(\mathbf{x}_{\mathcal{A}})$ is the vector consisting of $n_{i,a}(\mathbf{x}_{\mathcal{A}})$ for all i and a and $\mathbf{e}(\mathbf{x}_{\mathcal{A}})$ is the vector consisting of $e_a(\mathbf{x}_{\mathcal{A}})$ for all a. For brevity, we suppress the dependence of $\mathbf{x}_{\mathcal{A}}$ where it is unambiguous. Then, as we will explain below, θ is defined as

$$\theta(\mathbf{n}, \mathbf{e}) := \frac{1}{2} \sum_{i \neq j} \sum_{a \neq b} n_{i,a} \cdot n_{j,b} + \sum_a e_a \left(1 - \frac{n_a}{n_{\mathcal{V}}}\right) \tag{A.2}$$

Here, i and j range over all class indices, and a and b range over all possible outcomes of test t. The first term on the right-hand side counts the number of edges that will be cut by selecting test t no matter what the outcome of t is. Such edges consist of hypotheses that disagree with each other at t and, as with all edges, lie in different classes. The second term counts the expected number of edges cut by t consisting of hypotheses that agree with each other at t. Such edges will be cut by t iff they disagree with h^* at t. The edges $\{h, h'\}$ with $h, h' \in \mathcal{V}(\mathbf{x}_\mathcal{A})$ and $P(X_t = a \mid h) = P(X_t = a \mid h') = 1$ (of which there are e_a) will be cut by t iff $X_t \neq a$. Since we assume a uniform prior, $\mathbb{P}[X_t \neq a \mid \mathbf{x}_\mathcal{A}] = 1 - n_a/n_\mathcal{V}$ for any partial realization $\mathbf{x}_\mathcal{A}$ with $t \notin \mathcal{A}$, hence the expected contribution of these edges to $\Delta(t \mid \mathbf{x}_\mathcal{A})$ is $e_a(1 - n_a/n_\mathcal{V})$, from whence we get the second term.

Now fix $\mathbf{x}_\mathcal{B} \succ \mathbf{x}_\mathcal{A}$. Our strategy for proving $\Delta(t \mid \mathbf{x}_\mathcal{B}) \leq \Delta(t \mid \mathbf{x}_\mathcal{A})$ is as follows. As more observations are made, the version space can only shrink, i.e. $\mathcal{V}(\mathbf{x}_\mathcal{B}) \subseteq \mathcal{V}(\mathbf{x}_\mathcal{A})$. This means that for all i and a, $n_{i,a}$ is nonincreasing, i.e., $n_{i,a}(\mathbf{x}_\mathcal{B}) \leq n_{i,a}(\mathbf{x}_\mathcal{A})$. Note we may interpret e_a as a function of the variables in $\{n_{i,a} : 1 \leq i \leq m, a \in \mathcal{X}\}$, and that it is nondecreasing in each $n_{i,a}$, so we may also deduce that $e_a(\mathbf{x}_\mathcal{B}) \leq e_a(\mathbf{x}_\mathcal{A})$ for all a. Hence we consider a parameterized path $p(\tau)$ in $\mathbb{R}^{(m+1)\cdot\ell}$ from $p(0) := (\mathbf{n}(\mathbf{x}_\mathcal{B}), \mathbf{e}(\mathbf{x}_\mathcal{B}))$ to $p(1) := (\mathbf{n}(\mathbf{x}_\mathcal{A}), \mathbf{e}(\mathbf{x}_\mathcal{A}))$. Then by integrating along the path we obtain

$$\Delta(t \mid \mathbf{x}_\mathcal{A}) - \Delta(t \mid \mathbf{x}_\mathcal{B}) = \int_{\tau=0}^{1} \left(\frac{d(\theta \circ p)}{d\tau} \right) d\tau. \tag{A.3}$$

We require that at each point in p it holds that $e_a = \frac{1}{2} \sum_{i \neq j} \sum_{b \neq a} n_{i,b} \cdot n_{j,b}$ for all a, and also ensure that p is nondecreasing in each coordinate. There exists a path meeting these requirements, since $(\mathbf{n}(\mathbf{x}_\mathcal{B}), \mathbf{e}(\mathbf{x}_\mathcal{B})) \leq (\mathbf{n}(\mathbf{x}_\mathcal{A}), \mathbf{e}(\mathbf{x}_\mathcal{A}))$ and each e_a is nondecreasing in each $n_{i,b}$ variable. This implies $\partial n_{i,a}/\partial\tau \geq 0$ and $\partial e_a/\partial\tau \geq 0$ for all i and a. Hence we can prove the integral is nonnegative by applying the chain rule for the derivative to obtain

$$\frac{d(\theta \circ p)}{d\tau} = \sum_{i,a} \frac{\partial \theta}{\partial n_{i,a}} \frac{\partial n_{i,a}}{\partial \tau} + \sum_{a} \frac{\partial \theta}{\partial e_a} \frac{\partial e_a}{\partial \tau}$$

and then proving that $\partial\theta/\partial n_{i,a} \geq 0$ and $\partial\theta/\partial e_a \geq 0$ for all i and a. Next, observe that $\partial\theta/\partial e_a = (1 - n_a/n_\mathcal{V}) \geq 0$. So fix a class index k and an outcome c and consider $\partial\theta/\partial n_{k,c}$. Elementary calculus tells us that

$$\frac{\partial \theta}{\partial n_{k,c}} = \sum_{j \neq k,\, b \neq c} n_{j,b} + \sum_{b} \frac{e_b n_b}{n_\mathcal{V}^2} - \frac{e_c}{n_\mathcal{V}} \tag{A.4}$$

This quantity is nonnegative iff

$$n_\mathcal{V} e_c \leq n_\mathcal{V}^2 \cdot \sum_{j \neq k,\, b \neq c} n_{j,b} + \sum_{b} e_b n_b \tag{A.5}$$

Now substitute $\frac{1}{2} \sum_{i \neq j} \sum_{b \neq c} n_{i,b} \cdot n_{j,b}$ for e_c to obtain

$$n_\mathcal{V} e_c = \frac{n_\mathcal{V}}{2} \sum_{i \neq j} \sum_{b \neq c} n_{i,b} \cdot n_{j,b} \leq n_\mathcal{V} \cdot \left(\sum_{j \neq k,\, b \neq c} n_{j,b} \right) \left(\sum_{i,a} n_{i,a} \right) \leq n_\mathcal{V}^2 \cdot \left(\sum_{j \neq k,\, b \neq c} n_{j,b} \right) \tag{A.6}$$

Since $\sum_b e_b n_b \geq 0$, we obtain Eq. (A.5) from Eq. (A.6) by inspection, and hence $\partial\theta/\partial n_{k,c} \geq 0$ for all k and c. This completes the proof of the adaptive submodularity of f under a uniform prior.

We now show how to reduce the general prior case to the uniform prior case. Fix any prior P with rational probabilities, i.e. $P(h) \in \mathbb{Q}$ for all h. Then there exists $d \in \mathbb{N}$ and function $k : \mathcal{H} \to \mathbb{N}$ such that such that $P(h) = k(h)/d$. Create a new instance containing d hypotheses, where for each $h \in \mathcal{H}$ there are $k(h)$ copies of h, denoted by $h^1, \ldots, h^{k(h)}$. Each copy of h induces the same conditional distribution of test outcomes $P(X_1, \ldots, X_N \mid h)$. All copies of h belong to the same class, and copies of h and h' belong to the same class iff h and h' do. Finally, assign a uniform prior to this new instance. Then the adaptive submodularity of f on this new instance implies the adaptive submodularity on the original instance, if the weight of edge $\{h, h'\}$ in the original instance is proportional to the number of edges between the copies of h and the copies of h' in the new instance. That is, it suffices to set $w(\{h, h'\}) \propto k(h) \cdot k(h')$, and our choice of weight function, $w(\{h, h'\}) := P(h) \cdot P(h')$, satisfies this condition. $\qquad\square$

B A Bad Example for the Info-Gain and Value of Information Criteria

A popular heuristic for the Optimal Decision Tree problem are to adaptively greedily select the test that maximizes the *information gain* in the distribution over hypotheses, conditioned on all previous test outcomes. The same heuristic can be applied to the Equivalence Class Determination problem, in which we compute the information gain with respect to the entropy of the distribution over *classes* rather than hypotheses. Let π_{IG} denote the resulting policy for Equivalence Class Determination.

Another common heuristic for Optimal Decision Tree is to adaptively greedily select the test maximizing the *Bayesian decision-theoretic value of information* (VoI) criterion. Recall the value of information of a test t is the expected reduction in the expected risk of the minimum risk decision, where the risk is the expected loss. Formally, consider the Bayesian decision-theoretic setup described in §4. The VoI criterion myopically selects test to maximize

$$\Delta_{\text{VoI}}(t \mid \mathbf{x}_{\mathcal{A}}) := \min_d \mathbb{E}_H[\ell(d, H) \mid \mathbf{x}_{\mathcal{A}}] - \mathbb{E}_{x_t \sim X_t \mid \mathbf{x}_{\mathcal{A}}}\left[\min_d \mathbb{E}_H[\ell(d, H) \mid \mathbf{x}_{\mathcal{A}}, x_t]\right].$$

This heuristic can be also be applied to the Equivalence Class Determination problem, by taking the decision set $\mathcal{D}$ to be the set of equivalence classes, and the loss function to be the 0–1 classification loss function, i.e., $\ell(d, H) = \mathbf{1}[H \notin d]$. Let π_{VoI} denote the resulting policy.

In this section we present a family of Equivalence Class Determination instances for which both π_{IG} and π_{VoI} perform significantly worse than the optimal policy.

Theorem 7. *There exists a family of Equivalence Class Determination instances with uniform priors such that $c(\pi_{IG}) = \Omega\left(n/\log(n)\right) c(\pi^*)$ and $c(\pi_{VoI}) = \Omega\left(n/\log(n)\right) c(\pi^*)$, where n is the number of hypotheses and π^* is an optimal policy.*

In fact, we will prove a lower bound for each policy within a large family of adaptive greedy policies which contains π_{IG} and π_{VoI}, which we call *posterior–based*. Informally, this family consists of all greedily policies that use only information about the posterior equivalence class distribution to select the next test. More precisely, these policies define a potential function Φ which maps distributions of distributions over equivalence classes to real numbers, and at each time step select the test t which maximizes the Φ of the posterior distribution (over test outcomes x_t) of the posterior distribution over equivalence classes generated by adding x_t to the previously seen test outcomes. In the event of a tie, we select any test maximizing this quantity at random. The information gain policy is posterior–based; Φ is simply -1 times the expected entropy of the posterior equivalence-class distribution. Likewise, the value of information policy is also posterior–based; Φ is simply -1 times the expected loss of the best action for the posterior equivalence-class distribution. Hence to prove Theorem 7 it suffices to prove the following more general theorem.

Theorem 8. *There exists a family of Equivalence Class Determination instances with uniform priors such that $c(\pi) = \Omega\left(n/\log(n)\right) c(\pi^*)$ for any posterior–based policy π, where n is the number of hypotheses and π^* is an optimal policy.*

Proof. Fix integer parameter $q \geq 1$. There are $m = 2^q$ classes $\mathcal{H}_a$ for each $1 \leq a \leq 2^q$. Each $\mathcal{H}_a$ consists of two hypotheses, $h_{a,0}$ and $h_{a,1}$. We call a the *index* of $\mathcal{H}_a$. The prior is uniform over the hypotheses $\mathcal{H} = \{h_{a,v} : 1 \leq a \leq m, 0 \leq v \leq 1\}$. There are four types of tests, all with binary outcomes and all of unit cost. There is only one test of the first type, t_0, which tells us the value of v in the realized hypothesis $h^*_{a,v}$. Hence for all a, $H = h_{a,v} \Rightarrow X_{t_0} = v$. Tests of the second type are designed to help us quickly discover the index of the realized class via binary search if we have already run t_0, but to offer no information gain whatsoever if t_0 has not yet been run. There is one such test t_k for all t with $1 \leq k \leq q$. For $z \in \mathbb{N}$, let $\phi_k(z)$ denote the kth least-significant bit of the binary encoding of z, so that $z = \sum_{k=1}^{\infty} 2^{k-1}\phi_k(z)$. Then for each $h_{a,v}$ we have $H = h_{a,v} \Rightarrow X_{t_k} = \mathbf{1}[\phi_k(a) = v]$. Tests of the third type are designed to allow us to do a (comparatively slow) sequential search on the index of the realized class. Specifically, we have tests t_k^{seq} for all $1 \leq k \leq m$, such that $H = h_{a,v} \Rightarrow X_{t_k^{\text{seq}}} = \mathbf{1}[a = k]$. Finally, tests of the fourth type, $\left\{t_k^{\text{dumb}} : k \in \mathbb{N}\right\}$, are dummy tests that reveal no information at all. Formally, $X_{t_k^{\text{dumb}}}$ always equals zero.

Given this input, suppose $H = h_{a,v}$. One solution is to run t_0 to find v, then run tests $t_1, \ldots, t_q$ to determine $\phi_k(a)$ for all $1 \le k \le q$ and hence to determine a. This reveals the value of H, and hence the class H belongs to. Since the tests have unit cost, this policy π' has cost $c(\pi') = q + 1$.

Next, fix a posterior–based policy π and consider what it will do. Call a class *possible* if not all of its hypotheses have been ruled out by tests performed so far. Note that all possible classes contain the same number of hypotheses, because they initially have two, and each test t_k that can reduce the size of a possible class to one, will reduce the size of every possible class to one. This, and the fact that the prior is uniform, implies that the posterior equivalence-class distribution is uniform over the remaining possible classes. If no tests in $\{t_k : 0 \le k \le q\}$ have been run, as is initially the case, any single test in this set will not change the posterior equivalence-class distribution. Hence, as measured with respect to Φ, such tests are precisely as good as the dummy tests. If these tests are each better than any test in $\{t_k^{\text{seq}} : 1 \le k \le m\}$, then π selects among $\{t_k : 0 \le k \le q\} \cup \{t_k^{\text{dumb}} : k \in \mathbb{N}\}$ at random. Since there are infinitely many dummy tests, with probability one a dummy test is selected. Since the posterior remains the same, π will repeatedly select a test at random from this set, resulting in an infinite loop as dummy tests are selected repeatedly *ad infinitum*. Otherwise, some test t_k^{seq} is preferable to the other tests, measured with respect to Φ. In the likely event that t is not the index of H, we are left with a residual problem in which tests in $\{t_k : 0 \le k \le q\}$ still have no effect on the posterior, there is one less class, and the prior is again uniform. Hence our previous argument still applies, and π will either enter an infinite loop or will repeatedly select tests in $\{t_k^{\text{seq}} : 1 \le k \le m\}$ until a test has an outcome of 1. Thus in expectation π costs at least $c(\pi) \ge \frac{1}{m} \sum_{z=1}^{m} z = (m+1)/2$. Since $m = 2^q$, $n = 2m$, and $c(\pi^*) \le c(\pi') = q + 1 = \log_2(n)$ we infer

$$c(\pi) \ge \frac{m}{2} = \left(\frac{n}{4 \log_2(n)} \right) c(\pi^*)$$

which completes the proof. $\qquad\square$